Towards Distribution-shift Robust Text Classification of Emotional Content

Luana Bulla
Institute of Science and Technology of Cognition, National Research Council
luana.bulla@istc.cnr.it

Aldo Gangemi
University of Bologna
aldo.gangemi@unibo.it

Misael Mongiovì
Institute of Science and Technology of Cognition, National Research Council
misael.mongiovi@istc.cnr.it

Abstract

Supervised models based on Transformers have been shown to achieve impressive performances in many natural language processing tasks. However, besides requiring a large amount of costly manually annotated data, supervised models tend to adapt to the characteristics of the training dataset, which are usually created ad-hoc and whose data distribution often differs from the one in real applications, showing significant performance degradation in real-world scenarios. We perform an extensive assessment of the out-of-distribution performances of supervised models for classification in the emotion and hate-speech detection tasks and show that NLI-based zero-shot models often outperform them, making task-specific annotation useless when the characteristics of final-user data are not known in advance. To benefit from both supervised and zero-shot approaches, we propose to fine-tune an NLI-based model on the task-specific dataset. The resulting model often outperforms all available supervised models both in distribution and out of distribution, with only a few thousand training samples.

1 Introduction

Supervised text classification based on Transformers has recently achieved considerable performances, benefiting many applications in social (Mozafari et al., 2020) technological (Callaghan et al., 2021) and biomedical (Jin and Szolovits, 2020) domains, just to mention a few. However, these systems rely on large amounts of manually annotated data that are often expensive to obtain. Furthermore, to guarantee reasonable performances, supervised systems need to be trained on data that have the same distribution as the one in the deployed scenario (Koh et al., 2021). This requires a careful choice of data to annotate that is sometimes impossible to achieve because of the difficulty to infer in advance the characteristics of runtime data, and considering the potential evolution of data features during the system's lifetime (D'Amour et al., 2020). Recent work has shown that Transformers are more robust than other machine learning models to change in domain and distribution (Hendrycks et al., 2020). However, the decrease in performance due to the distribution shift is still a major issue of supervised models (Yang et al., 2022b).

Figure 1 shows the degradation in performances of models when applied to a different distribution. We consider three emotion classification tasks (with different taxonomies) and a hate speech detection task, and report in-distribution (ID) performances, when the model is validated on the same dataset (light blue bars), in comparison with out-of-distribution (OOD) performances, i.e. validated on different datasets (dark red bars). The drop in performance is significant, often overcoming 30% and sometimes reaching almost 50%. This makes models trained on certain data barely generalizable to other data, limiting drastically their scope.

Recent zero-shot models (Yin et al., 2019; Liu et al., 2021) have gained popularity thanks to their ability to reduce the dependency on task-specific annotated data by enabling models to predict previously unseen labels. For instance, models trained for Next Sentence Prediction (NLP) or Natural Language Inference (NLI) tasks can be applied to infer

Findings of the Association for Computational Linguistics: ACL 2023, pages 8256–8268

whether a certain textual label is associated with a sentence (Yin et al., 2019). Although supervised task-specific trained models typically outperform zero-shot approaches in the training dataset, it is reasonable to question how they compare when the supervised approach is trained on a different dataset.

In this work, we make a comprehensive assessment of the OOD performances of emotion detection and hate speech detection models in comparison with a NLI-based zero-shot model. Surprisingly, our results show that the zero-shot approach almost always outperforms the supervised models, suggesting that labeling a large amount of data is not beneficial when the data distribution is not a-priori known. To take advantage of both approaches we propose to adapt and fine-tune a NLI model with task-specific data. We show that a small amount of training data is sufficient to achieve performances that are often superior to the top-performing supervised models available, either ID or OOD.

Our contribution can be summarized as follows: (1) we perform a comprehensive assessment of the OOD performance of supervised models for classification (multi-class, multi-label, binary) of emotive content in comparison to an NLI-based approach that does not require specific training, and we show that the latter often achieves higher performance; (2) we propose fine-tuning an NLI model on task-specific data and show experimentally that this solution achieves competitive performances both ID and OOD with only a few thousand samples; (3) we extensively discuss our results and give useful indications for achieving significant ID and OOD performances with a small annotation cost.

2 Related Works

Developing models that are robust to domain and distribution shift is one of the most intriguing yet challenging tasks in various machine learning applications (Koh et al., 2021) including computer vision (Ibrahim et al., 2022; Yang et al., 2022a; Larson et al., 2022) and NLP (Csordás et al., 2021; Malinin et al., 2021; Hendrycks et al., 2020). We refer to Zhou et al. (Zhou et al., 2022) for an extensive survey on domain generalization. While some works offer a more theoretical perspective on the topic (Arora et al., 2021a; Ren et al., 2019), general work in the NLP field has been focused mainly on developing benchmarks for evaluating the out-of-distribution robustness of models. Hendrycks et al. (Hendrycks et al., 2020) studies OOD generalization for seven NLP datasets in the tasks of sentiment classification, semantic similarity, reading comprehension, and textual implication and show that pre-trained Transformers adapt better to OOD data. Yang et al. (Yang et al., 2022b) propose a unified benchmark called GLUE-X to evaluate OOD robustness in NLP systems. They collect 13 datasets covering tasks such as sentiment analysis, natural language inference, sentence pair similarity, textual similarity, and linguistic acceptability. For each task, they select a dataset for training and other datasets for OOD evaluation. The study shows that better OOD accuracy is needed for NLP tasks, due to the noticeable loss of performance with respect to the ID settings. Both works do not compare the performances to zero-shot approaches and do not propose specific methods for increasing OOD robustness. Furthermore, they do not consider the tasks of emotion detection and hate-speech detection.

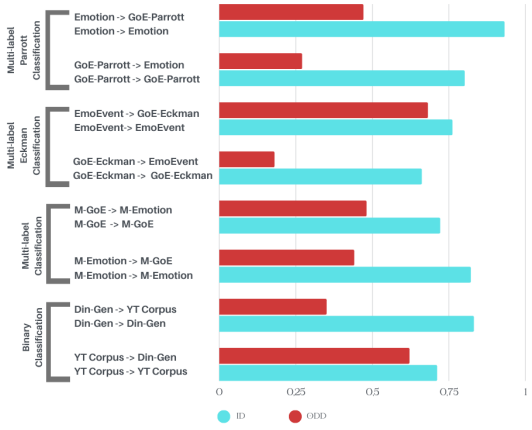

Figure 1: Performance degradation (as weighted F1-score) of supervised models when applied to a different distribution (dataset). Light blue bars represent ID performances, while dark red bars show OOD results.

Approaches to deal with the distribution shift problem include OOD detection (Arora et al., 2021b) and Mixture of Experts (MoE) models (Guo et al., 2018). OOD detection aims at recognizing OOD text to give awareness of the potential degradation in performances, while MoE models tend to combine domain-specific models to improve performances in multi-domain contexts. Both these approaches are out of the scope of our work since they do not specifically focus on assessing and improving the performances of models over unseen domains and data distribution changes.

Specific studies on text classification related to ours usually focus on domain generalization by

training on text in one domain and testing on a different domain within the same dataset. Although related, these approaches do not consider domain-independent differences that occur across datasets concerning e.g. text features (e.g. length), linguistics features (e.g. use of slang) and annotation processes. PADA (Ben-David et al., 2022) generates domain-related features and adds them to the text to enable the model to adapt to different domains. Other studies refer to specific tasks such as moral value classification (Liscio et al., 2022) and sentiment analysis (Fu and Liu, 2022; Zhang et al., 2022a; Li et al., 2022; Luo et al., 2022; Liu and Zhao, 2022). Despite not considering the model generalization across datasets, and being often application-specific, these methods do not make any assessment with zero-shot learning nor consider building upon them to improve OOD performance.

To the best of our knowledge, the only studies on distribution shift that consider the emotion and hate-speech detection tasks are the work of Toraman et al. (Toraman et al., 2022), which evaluates how BERT generalizes across abusive language detection datasets, and Zeng et al. (Zeng et al., 2022) that propose a CNN-based broad learning model for cross-domain emotion classification. The first study on abusive language detection does not compare with zero-shot models nor proposes a method for OOD generalization. The Zeng et al. work considers a multi-domain dataset obtained by collecting data from Chinese E-commerce platforms and performs the assessment across domains. Again they do not perform a comparison with zero-shot models nor evaluation across datasets.

Another line of research related to our work concerns zero-shot and prompt-based models. Pushp et al. (Pushp and Srivastava, 2017) propose and evaluate three LSTM architectures for zero-shot classification that combine text embedding with label embedding to determine whether the label is related to the input text. Yin et al. (Yin et al., 2019) provide datasets, a standard for evaluation and state-of-the-art baselines for zero-shot classification. Barker et al. (Barker et al., 2021) propose performing supervised classification on known labels, then applying NLI for cases that do not qualify for previously known labels. Zhang et al. (Zhang et al., 2022b) propose a meta-learning framework to learn to calibrate the class and sample representations from both seen and virtual unseen classes. Other studies focus on the impact of different prompts on performances (Liu et al., 2021). In particular, we highlight the work of Plaza-del-Arco et al. (Plaza-del Arco et al., 2022) which compares different ways to build the hypothesis prompt for NLI-based emotion detection. Although we adopt an NLI-based zero-shot model in our work, taking inspiration from the work of Yin et al. (Yin et al., 2019) and Plaza-del-Arco et al. (Plaza-del Arco et al., 2022), no other work that we are aware of makes an extensive comparison of OOD performances of supervised models with zero-shot models and fine-tuning of the latter, finding the sweet spot between no-specific training and a fully supervised approach.

To the best of our knowledge, there are no extensive state-of-the-art studies focusing on the OOD robustness (across datasets) of supervised models in emotion classification tasks. In general, we are not aware of any study that compares OOD performances of supervised models for text classification with zero-shot models and assesses the best way to fine-tune a zero-shot model.

3 Materials and Methods

We describe in detail the benchmark data and models of our work. In Section 3.1 we discuss datasets. Section 3.2 focuses on the analysis of the state-of-the-art supervised models and the NLI-based systems we employ. Table 1 summarizes the material of our study, including classification tasks, taxonomies, datasets and supervised models.

3.1 Datasets

We conduct our experimental study on ten datasets for multi-class, multi-label and binary classification. Specifically, we focus on datasets for emotion and hate speech detection.

For the multi-class emotion classification task, we apply five distinct benchmarks and study two different taxonomies. The first set includes the range of the Primary Emotions of Parrott theory (Parrott, 2001) (i.e. "love", "joy", "sadness", "anger", " fear", " and "surprise") and is covered by Emotion corpus (Saravia et al., 2018) and a scaled-down version of the GoEmotion dataset (Demszky et al., 2020) (GoE-Parrott). We consider GoEmotion for its wide range of content, labels and data qualities, which make it suitable for fitting emotion taxonomies of other datasets used in our study. For this reason, we generate two additional customized versions of this benchmark. The first one (GoE-Ekman) is designed for multi-class detection based

Task	Typology	Taxonomy	Datasets	Models
Parrott Emotion	Multi-class	Love, joy, sadness, anger, fear, surprise	GoE-Parrott Emotion (Saravia et al., 2018)	E-T5 E-Bert GoE-Bert
Ekman Emotion	Multi-class	Disgust, joy, sadness, anger, fear, surprise and neutral	GoE-Ekman EmoEvent (Plaza-del Arco et al., 2020) XED ("Ohman et al., 2020)	E-BERTweet (Pérez et al., 2021) E-DistilRoBERTa (Hartmann, 2022) Emo-Bert
Multi Emotion	Multi-label	Disgust, joy, sadness, anger, fear, love, optimism and surprise	M-GoE M-Emotion (Mohammad et al., 2018)	M-Bert M-GoE Bert
Binary-HS	Binary	Hate, Not Hate	Din-Gen (Vidgen et al., 2020) YouTube (Ljubešić et al., 2021) WSF-HS (De Gibert et al., 2018)	LFTW-RoBERTa (Vidgen et al., 2020) YT-Bert (Ljubešić et al., 2021)

Table 1: Overview of the material of our experimental study. Supervised models, taxonomies and datasets are grouped according to the type of classification they apply to.

on Ekman's theory of emotions (Ekman, 1992) ("disgust", "joy", "sadness", "anger", "fear" and "surprise", plus an additional "neutral" label) and adapts to the XED dataset ("Ohman et al., 2020) and the tweet-based EmoEvent corpus (Plaza-del Arco et al., 2020). The second one (M-GoE), allows us to fit the M-Emotion corpus (Mohammad et al., 2018), a tweet-based restricted dataset for multi-label classification. By taking all emotions that overlap between GoEmotion and M-Emotion, we obtain a third taxonomy based on eight labels (i.e. "disgust", "joy", "sadness", "anger", "fear", "love", "optimism" and "surprise"). In the second stage, we focus on the binary hate-speech detection task. In this scenario, we employ the Dynamically Generated dataset (Vidgen et al., 2020) (Din-Gen), the YouTube HS corpus (Ljubešić et al., 2021) (YouTube), and the WSF-HS dataset (De Gibert et al., 2018). The former is built by an iterative annotation process, starting from a collection of previously released hate speech datasets, the second is composed of YouTube comments captured during the time of the COVID-19 pandemic, while the third focuses on a random collection of offensive forum posts. Further details on the employed datasets are given in the supplemental material.

3.2 Reference Models

We employ a group of supervised models designed to address multi-class, multi-label, and binary classification, in order to evaluate their OOD performances, i.e. their performances on a different dataset than the one used for training. As a comparison, we examine the results of three alternative NLI-based system configurations, seeing how unsupervised models perform in this context. The following paragraphs provide further information on the first and second groups (Sect. 3.2.1, 3.2.2).

3.2.1 Supervised Models

We use eight models for emotion detection, six of which are focused on a multi-class classification scenario and the other two ones on multi-label classification. To perform an ODD evaluation on all datasets available and since no trained model is suited to some of them, we trained four standard BERT classifiers on the missing datasets (i.e. GoE-Parrott, M-GoE, and M-Emotion) obtaining checkpoints that we name GoE-Bert, M-GoE Bert, M-Bert. The classifiers employ the pre-trained BERT-base checkpoint and apply a dropout layer, a linear layer and then a softmax on the pooled output embedding of the CLS token. We also train the same BERT-based architecture on the EmoEvent dataset (Plaza-del Arco et al., 2020) (Emo-Bert) to compare it to BERTweet, which has been pre-trained on tweet data. The tune of hyperparameters was conducted on the validation set through grid search taking into consideration a range of parameters ranging from 0.1 to 0.4 for the dropout, among 10^{-5}, $3 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ for the learning rate, and between 32 and 64 for the batch size. The number of epochs was set to 10. For each configuration, we performed a single run. For the multi-class classification task, we also employ the BERT-based E-Bert model[1] and the T5-based (E-T5) system[2]. Both of them are trained on the Emotion dataset (Saravia et al., 2018) and explore the Parrott theory perspective. From Ekman's taxonomy, we consider the RoBERTa-based E-BERTweet (Pérez et al., 2021), and E-DistilRoBERTa (Hartmann, 2022) models, which are trained on the EmoEvent corpus (Plaza-del Arco et al., 2020) and on a combination of six emotional datasets (Hartmann, 2022), respectively.

For binary hate-speech detection, we employ YT-Bert (Ljubešić et al., 2021) and LFTW-

[1] huggingface.co/bert-base-uncased-emotion
[2] huggingface.co/t5-base-finetuned-emotion

RoBERTa (Vidgen et al., 2020). The former has been trained on the YouTube corpus (Ljubešić et al., 2021) while the latter refers to the Din-Gen dataset (Vidgen et al., 2020).

3.2.2 NLI-based classifiers

Inspired by the work of Yin et al. (Yin et al., 2019), we employ pre-trained NLI models as ready-made zero-shot sequence classifiers. We create a hypothesis for each potential label and use the input text as an NLI premise. For the hypothesis construction, we use the prompt "This sentence expresses <label>". We use "discrimination and hate" as label for hate speech. Different prompts are also employed for the prompt analysis in Section 4.4. To determine which emotion is prevalent in the input text from an NLI perspective we take the emotion that corresponds to the highest-scoring entailment output. To manage neutrality and for binary classification, we apply a 0.5 cut-off on the normalized entailment score. For multi-label classification, we take all emotions that correspond to a normalized entailment score above or attained to 0.5. Since there is no specific training phase, the approach is particularly useful when there are no high-quality task-specific annotated samples. Furthermore, the method is applicable to a variety of document types in different domains.

We consider three checkpoints as NLI models: MNLI-Bart-large[3], MNLI-RoBERTa-large [4] and MNLI-DeBERTa-large[5], all trained on the MultiNLI (MNLI) dataset (Williams et al., 2017). For more details on the different configurations of the NLI models on the taxonomies and datasets examined (Sect. 3.1), we refer to Sect. 4.

3.2.3 Fine-tuning NLI-based classifiers

We propose optimizing NLI models (we take MNLI-RoBERTa-large as reference) on task-specific datasets to take advantage of both zero-shot and supervised methods. We replace the last linear layer of the NLI model to fit the classification taxonomy. The resulting architecture is fine-tuned on the target dataset. During fine-tuning the parameters of the last linear classification layer are learned from scratch, while the remaining parameters are tuned. The tune of hyperparameters was conducted on the validation set through grid search taking into consideration a range of parameters ranging from 0.1 to 0.4 for the dropout, among 10^{-5}, $3 \cdot 10^{-5}$ and $5 \cdot 10^{-5}$ for the learning rate, and between 32 and 64 for the batch size. The number of epochs was set to 10. For each configuration, we performed a single run.

[3] huggingface.co/facebook/Mnli-Bart-large
[4] huggingface.co/roberta-large-mnli
[5] huggingface.co/deberta-large-mnli

4 Results and Evaluation

We assess the OOD performances of supervised models in comparison with NLI-based classification. We group our evidence by looking at three main classification problems: multi-class (where an item can be associated with only one label), multi-label (where an item can be associated with more than one label) and binary. We set out to investigate performances on emotion-domain-specific detection tasks as a unifying framework across all experiments. We present the details of our experimental settings and the results in Sections 4.1, 4.2 and 4.3. All output data are available on GitHub[6]. We also evaluate the performances of different prompts in Section 4.4. In the last parts of our experimental analysis we evaluate the NLI-with-fine-tuning method discussed in Section 3.2.3 at varying the number of training samples and in comparison with fully supervised systems (Sect. 4.5). Unless differently specified, all performances reported refer to the F1-score. For multi-class and multi-label classification, we consider the weighted F1. All experiments have been run on a server with 2 CPU Intel Xeon Gold 6238R 2.20GHz with 640GB RAM and two GPU A100 40GB. As a rough estimation, our experiments took in total about 30 GPU days.

4.1 Multi-class classification

We examine the models' performances in multi-class emotion detection considering two different taxonomies. The former, based on the Parrott theory, considers six different emotions (joy, love, sadness, surprise, anger and fear), while the latter focuses on Ekman's theory with the addition of a seventh category for expressions devoid of emotional connotations. By expanding the taxonomic coverage, we intend to test the models' ability to discriminate between more or less semantically complex labels in uncorrelated datasets.

In the first scenario, we compare OOD performances of supervised models E-T5[7], E-Bert[8], and GoE-Bert[9], discussed in Sect. 3.2.1, with

[6] https://github.com/LuanaBulla/Text-Classification-of-Emotional-Content
[7] huggingface.co/t5-base-finetuned-emotion
[8] huggingface.co/bert-base-uncased-emotion
[9] link hidden for blind review

Models	GoE-Parrott	Emotion
E-T5	0.51	-
E-Bert	0.47	-
GoE-Bert	-	0.27
MNLI-BART-large	0.63	0.51
MNLI-RoBERTa	0.72	0.52
MNLI-DeBERTa	**0.74**	**0.54**

Table 2: F1-score for each supervised and NLI-based model on the multi-class emotion detection task, by adopting the first emotion taxonomy (i.e. joy, love, sadness, surprise, anger and fear). The table shows the OOD performances of the models. Cells with the hyphen indicate that the training dataset is the same as the one shown in the column.

Models	GoE-Ekman	EmoEvent	XED
E-BERTweet	0.68	-	**0.47**
Emo-Bert	0.65	-	0.42
E-DistilRoBERTa	-	0.18	**0.47**
MNLI-BART-large	0.64	0.44	0.39
MNLI-RoBERTa	**0.74**	0.49	0.42
MNLI-DeBERTa	0.66	**0.53**	0.45

Table 3: F1-score for each supervised and NLI-based model and on the multi-class emotion detection task by adopting Ekman's emotion taxonomy. The table shows the OOD performances. The hyphen indicates cells where the evaluation dataset has been used in the training phase.

the NLI-based systems discussed in Sect. 3.2.2 (i.e. MNLI-Bart-large, MNLI-RoBERTa-large and MNLI-DeBERTa-large) on GoE-Parrot and Emotion datasets discussed in Sect. 3.1. The results (Table 2) show that MNLI-DeBERTa performs better in both cases, with an F1-score of 0.74 on GoE-Parrott and 0.54 on Emotions. Moreover, NLI-based systems always outperform supervised systems by a wide margin.

In the second scenario, we consider the EmoEvent corpus, the GoE-Ekman dataset and the XED dataset (discussed in Sect. 3.1) as benchmarks. The supervised models are E-BERTweet, E-DistilRoBERTa, Emo-Bert (all discussed in 3.2). Table 3, shows that the top-performing system is NLI-based on two over three cases. On EmoEvent every NLI-based system outperforms the supervised model by a wide margin. On XED, results are comparable (0.45 for NLI-based vs. 0.47 for supervised models). In this dataset, all models show suboptimal performances, which might indicate a lower quality of the data.

Models	M-Emotion	M-GoE
Multi-E Bert	-	0.48
M-GoE Bert	0.44	-
MNLI-BART-large	0.45	0.53
MNLI-RoBERTa	0.46	0.58
MNLI-DeBERTa	**0.49**	**0.63**

Table 4: F1-score for each supervised and NLI-based model in the multi-label emotion detection task. The table shows the OOD performances. Hyphens indicate cells where the training and test datasets correspond.

4.2 Multi-label classification

To evaluate the performance of models in a multi-label emotion scenario, we adopt a seven-base taxonomy - joy, disgust, love, optimism, sadness, surprise, anger and fear - and test on the M-Emotion and M-GoE datasets (Sect. 3.1). We train M-Bert and M-GoE Bert (Sect. 3.2) on the above corpora to compare supervised vs. NLI-based models. As shown in Table 4, MNLI-DeBERTa achieves the best performances in both M-Emotion and M-GoE datasets, with an F1-score of 49% and 63%, respectively. Again all NLI-based models always outperform supervised models.

4.3 Binary classification

In order to assess how models react to the data-shift problem in a binary classification context, we test their ability to detect hate speech from datasets that are not included in their training phase. Results are reported in Table 5. We use as a benchmark the datasets Din-Gen, YouTubeand WSF-HS, discussed in Sect. 3.1. The supervised models were trained on Din-Gen and YouTube, respectively. We compare the performances of the supervised models (LFTW-RoBERTa and YT-Bert) with NLI-based architectures (i.e. MNLI-Bart-large, MNLI-RoBERTa, MNLI-DeBERTa), presenting the outcome again in terms of weighted F1-score for each model. As shown in Table 5, on Din-Gen and YouTube all NLI-based classifiers outperform supervised models, with the top performances achieved by MNLI-DeBERTa and MNLI-RoBERTa, respectively, with an F1-score of 0.72 and 0.62. On the WSF-HS dataset, LFTW-RoBERTa achieves the best performances with a 67% F1-score. The good performances of LFTW-RoBERTa suggest that WSF-HS data have characteristics in common with Din-Gen (training dataset for LFTW-RoBERTa). This explanation is supported by results reported in Sect. 4.5, where the NLI model fine-tuned on Din-Gen is shown to sig-

Models	Din-Gen	YouTube	WSF-HS
LFTW-RoBERTa	-	0.35	**0.67**
YT-Bert	0.62	-	0.45
MNLI-BART-large	0.70	0.59	0.39
MNLI-RoBERTa	0.68	**0.62**	0.38
MNLI-DeBERTa	**0.72**	0.54	0.40

Table 5: F1-score for each supervised and NLI model on the binary Hate-Speech classification task. The table shows the OOD performances of the models. Again, the hyphen indicates cells where the training dataset corresponds to the evaluation dataset in the column.

nificantly outperform LFTW-RoBERTa. Moreover, in this dataset, NLI-based methods perform worse than in other datasets, probably due to a more prominent imbalance among hate and non-hate content (only 12% of hate content), with respect to the other datasets.

4.4 Prompt analysis

To assess the variability of NLI-based classifiers with different prompts, we performed a comprehensive prompt configuration analysis on all NLI-based systems taken into consideration in our study. We consider the prompt as a hypothesis for NLI that is given to evaluate its degree of entailment by the text item, considered as a premise. We consider three different settings, each linked to a separate prompt. In the first case, we emphasize a factual point of view that explicitly uses the content of the sentence as the subject (Prompt 1: "This sentence expresses <label>" - we use "discrimination and hate" as a label for hate speech). In the second instance, we use the label provided by the taxonomy as is (Prompt 2: "<label>"). As a third option, we assume a more individualized track, which speaks directly to emotionality and subjectivity (Prompt 3: "I feel <label>" for emotion, "This is hateful content" for hate speech). In most cases, the performances of different prompts are similar (within 5%, with a few exceptions). Prompt 1 usually outperforms the other prompts. This is expected since Prompt 1 puts the focus on the content, while Prompt 2 is not well semantically connected with the text and Prompt 3 puts the focus outside of the content (the subject is "I"). Detailed results in terms of F1-score are given in the supplemental material.

4.5 Fine-tuned NLI analysis

Following the methodology detailed in Section 3.2.3, we fine-tune checkpoint MNLI-RoBerta-large on eight different dataset configurations to solve multi-class, multi-label and binary classification. Our analysis (details and tables in the supplemental material) shows that the NLI-fine-tuned system always outperforms supervised models on the same dataset (ID) when the whole dataset is employed for training, with only two exceptions, probably due to the presence of insufficient data (i.e. on M-Emotion) and the comparison to a model specifically pre-trained on the same kind of data (i.e. BERTweet, pre-trained on tweets). However, this adaptability has a disadvantage in terms of ODD results, which do not always show performances achieved by zero-shot architectures.

To find a good trade-off between ID and OOD performances, we scale down the training set and study how the model behaves as the training sample size increases. We start with a random sample of 100 items and expand it exponentially by doubling its size at each step until we reach the full size[10]. Figure 2 reports the results on the multi-class setting with the Parrott taxonomy. The model is trained on GoE-Parrot. The figure shows the trend of NLI fine-tuned at varying the number of training samples both ID (on GoE-Parrot itself) and OOD (on Emotion). We also report the ID and OOD performances of the native supervised model (GoE-Bert) and of NLI without fine-tuning (dashed lines). For small training samples, both ID and OOD performances degrade w.r.t. NLI without fine-tuning, since the last classification layer has not had enough data to adapt. When the training data increase, both ID and OOD performances rapidly rise. While ID performances always increase, OOD performances reach a plateau and then start to decrease. This suggests that the model is over-adapting to the specific dataset and hence became less generalizable to other datasets.

Not all cases show the same behavior: sometimes both ID and OOD performances always increase (for small datasets), and sometimes both reach a plateau. Table 6 gives a general performance overview of all datasets with a training size of 3200 items. The fine-tuned NLI-based classifier (MNLI-RoB-FineTuned) outperforms all native supervised systems (we report the top performer) in an OOD setting in 7 over 10 cases, with comparable results in the remaining cases, while achieving top ID performances on four over eight datasets. We note that in two over three cases of slightly worse OOD performance of MNLI-RoB-FineTuned, NLI

[10] we provide specific configuration details for each classification task in the supplement

Models	GoE Parrott	Emotion	GoE Ekman	EmoEvent	XED	M-Emotion	M-GoE	Din-Gen	YouTube	WSF HS
Top-Supervised (ID)	0.80	**0.94**	0.66	**0.76**	-	**0.72**	**0.82**	**0.83**	0.71	-
MNLI-RoB-FineTuned (ID)	**0.90**	0.91	**0.85**	0.60	-	0.49	**0.82**	0.79	**0.82**	-
Top-Supervised (ODD)	0.51	0.27	**0.68**	0.18	**0.47**	0.44	0.48	**0.62**	0.35	0.67
MNLI-RoB-FineTuned (ODD)	**0.57**	**0.56**	0.66	**0.48**	0.46	**0.45**	**0.61**	0.60	**0.65**	**0.84**
MNLI-RoBERTa	0.72	0.52	0.74	0.49	0.42	0.46	0.58	0.68	0.62	0.38

Table 6: F1-score for the NLI-based classifier fine tuned on 3200 training samples (MNLI-RoB-FineTuned) on both ID and OOD settings, in comparison with the top-performer native supervised approach (Top-Supervised). We also report the performances of the NLI-based classifier without fine-tuning (MNLI-RoBERTa). Underlined values in the last row indicate that MNLI-RoBERTa achieves the best OOD performances. ID performances for XED and WSF-HS are missing since we do not have a model trained on such datasets. For these two columns, OOD performances correspond to the best performances of models trained on the other two datasets

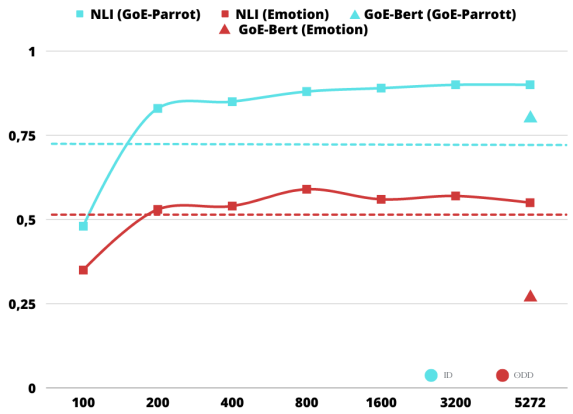

Figure 2: F1-score achieved by the NLI-based classifier with task-specific fine tuning on GoE-Parrot at varying the training sample size (x-axis) in comparison with the top-performer natively supervised system (GoE-Bert) and NLI-based without fine-tuning (dashed lines). Light blue color represents ID performances (on GoE-Parrot) while red color represents OOD performances (on Emotion).

without fine-tuning (MNLI-RoBERTa) achieves the best performance. In general, the zero-shot MNLI-RoBERTa model outperforms the refined NLI approach in the OOD scenario on five over ten datasets. However, in an ID setting, the former does not reach the performance of the latter. The fine-tuned approach with limited training data represents a good trade-off between ID and OOD performances. The complete plots for all datasets are available in the supplemental material.

4.6 Discussion

Our investigation compares supervised and NLI-zero-shot models with a focus on different typologies of emotion detection and hate speech recognition both in and out-of-distribution. This provides a comprehensive explanation of the limitations and advantages of both methodologies for the two scenarios. According to our results, the supervised models show good ID performance at the price of a significant drop in OOD performance. In contrast, unsupervised zero-shot systems excel in OOD settings but do not outperform supervised models in ID contexts. A reasonable compromise between the two methodologies is the NLI-fine-tuned method, which improves ODD results compared to supervised systems and achieves good performance compared to the zero-shot approach in an ID setting.

In a situation where limited training data are available, the fine-tuned NLI system has the advantage of achieving a good trade-off between ID and OOD performance, with less training data than supervised models. Using a zero-shot NLI-based system is preferable in situations where the final data distribution is unknown. Furthermore, it requires less implementation time and no training dataset.

Our experimental analysis is not without limitations. We focused on emotive content (emotion classification and hate speech detection), therefore our results might not be extendable to other domains. Emotions have a certain degree of subjectivity that can affect the annotation process by making data annotator-dependent. In other fields, this might not be the case. Moreover, our analysis is limited to ten datasets that we believe are representative of the work in this field. However, many other datasets are available, especially in the hate speech domain, and a wider evaluation might lead to a more definitive conclusion. Another limitation of our work is that we only considered NLI-based approaches as zero-shot models. Other zero-shot approaches might perform better, as pointed out by Ma et al. (Ma et al., 2021). In our experimental analysis, NLI-based approaches perform better than as reported in the Ma et al. work, and the difference might depend on implementation details. In any case, supposed better performances of other zero-shot approaches can only strengthen our conclusion, i.e. that it is possible to improve OOD

performances by limiting or completely avoiding task-specific training, which often requires a considerable annotation cost.

5 Conclusion

We made an extensive experimental analysis of the OOD performances of supervised Transformer-based classifiers trained on task-specific data with emotive content, in comparison with zero-shot approaches based on NLI that do not require specific training. Our results show that, although no-specific-training approaches are not able to perform as well as supervised models in the same dataset, they often achieve the best performance w.r.t. supervised models evaluated on a different dataset. We found that a mixed approach consisting in fine-tuning NLI-based classifiers with limited data reaches a good trade-off between ID and OOD performances.

Acknowledgement

We acknowledge financial support from the H2020 projects TAILOR: Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization – EC Grant Agreement number 952215 – and SPICE: Social Cohesion, Participation and Inclusion through Cultural Engagement – EC Grant Agreement number 870811, as well as from the Italian PNRR MUR project PE0000013-FAIR.

References

Udit Arora, William Huang, and He He. 2021a. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Udit Arora, William Huang, and He He. 2021b. Types of out-of-distribution texts and how to detect them. *arXiv preprint arXiv:2109.06827*.

Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. 2021. Ibm mnlp ie at case 2021 task 2: Nli reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 193–202.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics*, 10:414–433.

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J Brecha, et al. 2021. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature climate change*, 11(11):966–972.

Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. 2021. The devil is in the detail: Simple tricks improve systematic generalization of transformers. *arXiv preprint arXiv:2108.12284*.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*.

Paul Ekman. 1992. *Are there basic emotions?* American Psychological Association.

Yanping Fu and Yun Liu. 2022. Contrastive transformer based domain adaptation for multi-source cross-domain sentiment classification. *Knowledge-Based Systems*, 245:108649.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703.

Jochen Hartmann. 2022. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Mark Ibrahim, Quentin Garrido, Ari Morcos, and Diane Bouchacourt. 2022. The robustness limits of sota vision models to natural variation. *arXiv preprint arXiv:2210.13604*.

Di Jin and Peter Szolovits. 2020. Advancing pico element detection in biomedical text via deep neural networks. *Bioinformatics (Oxford, England)*, 36(12):3856–3862.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga,

Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Stefan Larson, Gordon Lim, Yutong Ai, David Kuang, and Kevin Leach. 2022. Evaluating out-of-distribution performance on document image classifiers. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tian Li, Xiang Chen, Zhen Dong, Weijiang Yu, Yijun Yan, Kurt Keutzer, and Shanghang Zhang. 2022. Domain-adaptive text classification with structured knowledge from unlabeled data. *arXiv preprint arXiv:2206.09591*.

Enrico Liscio, Alin Dondera, Andrei Geadau, Catholijn Jonker, and Pradeep Murukannaiah. 2022. Cross-domain classification of moral values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2727–2745.

Ning Liu and Jianhua Zhao. 2022. A bert-based aspect-level sentiment analysis algorithm for cross-domain text. *Computational Intelligence and Neuroscience*, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Nikola Ljubešić, Igor Mozetič, Matteo Cinelli, and Petra Kralj Novak. 2021. English YouTube hate speech corpus. Slovenian language resource repository CLARIN.SI.

Yun Luo, Fang Guo, Zihan Liu, and Yue Zhang. 2022. Mere contrastive learning for cross-domain sentiment analysis. *arXiv preprint arXiv:2208.08678*.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861.

Emily "Ohman, Marc P'amies, Kaisla Kajava, and J"org Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *The 28th International Conference on Computational Linguistics (COLING 2020)*.

W Gerrod Parrott. 2001. *Emotions in social psychology: Essential readings*. psychology press.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *arXiv preprint arXiv:1712.05972*.

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. 2021. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yılmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. 2022a. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022b. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Rong Zeng, Hongzhan Liu, Sancheng Peng, Lihong Cao, Aimin Yang, Chengqing Zong, and Guodong Zhou. 2022. Cnn-based broad learning for cross-domain emotion classification. *Tsinghua Science and Technology*, 28(2):360–369.

Kai Zhang, Qi Liu, Zhenya Huang, Mingyue Cheng, Kun Zhang, Mengdi Zhang, Wei Wu, and Enhong Chen. 2022a. Graph adaptive semantic transfer for cross-domain sentiment classification. *arXiv preprint arXiv:2205.08772*.

Yiwen Zhang, Caixia Yuan, Xiaojie Wang, Ziwei Bai, and Yongbin Liu. 2022b. Learn to adapt for generalized zero-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Sect. 4.6 Discussion we discuss the limitation of our work

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Our work mostly a comparison and improvement of text classification methods. It does not present potential risks

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Sect.1 summarize the paper's main claims. At the end of the introduction we describe the contribution of our work

☒ A4. Have you used AI writing assistants when working on this paper?
no, excluding the spell-checker integrated on overleaf

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 4 presents the experimental results

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We used previously proposed models and we refer to them for the parameters. We report in sect. 4 comutational budget and computing infrastructure

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We extensively discuss the experimental setup in sect. 3 and 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We specify in sect. 3 that all results refer to a single run

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Sect. 3.2 we give all details of our implementation including comprehensive references to all employed models, datasets and software

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.